Microblogging Short Text Classification based on Word2Vec

Yonghui Zhang[1, a*] and Jingang Liu[1, 2, b]

[1]Capital Normal University, Beijing 100048 China

[2]Institute of Computing Technology Chinese Academy of Sciences, Beijing 100089 China

[a]yonghuizhang2008@hotmail.com, [b]liujg2000@163.com

Keywords: Word2Vec; Features extension; Microblogging short text; SVM; Classification

Abstract. For the sparse features of the microblogging text, the author proposes a method of microblogging text classification based on the features extension by Word2Vec. We train the text by using Word2Vec tool and find the words which are similar to original features semantic as the features of short text. Then we expand the features to the original text, and finally classify the subject of microblogging text by using SVM method. Experimental results show that the method has high accuracy recall and F1 values compared with the traditional method of vector space model and LDA topic model.

Introduction

Text classification is to put a document into several predefined categories of one or several, and automatic classification of text is to use the vocabulary of rich features [1]. With the rapid development of the Internet and the popularity of Sina micro blogging and Ten cent microblogging and other popular social networking sites in the country, microblogging and other social media have become the platform of publishing and sharing the information thus accumulating a large amount of data. The data mining of is feasible, innovative and practical, and attracted widespread attention at home and abroad academia. The classification of microblogging text has an important role in in spam filtering, hot events discovery, network public opinion analysis, personalized recommendations, and other fields [2].

Currently, the classification for a longer text has some relatively mature technology, but the research on short text classification is in a stage of rapid development. The classification technology of long text is no longer suitable for short text due to the feature context-dependent and characteristic sparse. For classification of short text, mainly to take the help of external text [3-4] and external knowledge bases [5-6]. The former general use of the search engine results for short text feature extension, but the latter excavates the intrinsic link between the words within the text. The classification of microblogging has been made for some results in foreign. Bharath Sriram, who analyzed the author's Twitter profile and personal information related to the text, as microblogging feature space to a preliminary judgment on the data [7]. By analyzing the topic on Twitter over time to the classification on the short text that classification effect has been some improvement by Danesh Irani etc [8]. Domestic Chinese short text classification has also conducted a lot of research for the problem of lack for massive short text classification tagged corpus. Yuehong Cai, who proposed a classification algorithm based on attribute selection of semi-supervised short text [9]; Xiwei Wang, who used FP-Growth algorithm for mining co-occurrence relationship between the training set and test set feature item characteristics between items, and then introduced into the semantic information and improved ability to describe the formula HowNet DEF entry and then classified Chinese short text[10]; Chaozhen Lv, who used LDA topic model short text relating to extension, then classified the text[11]. The research of classification on Chinese microblogging is still in the exploratory stage.

Based on the above analysis, the author first uses the traditional vector space model for short text, and then trains the short text by using Word2Vec model, obtaining first n words with similar characteristics as the original part feature short texts, thus expanding short text feature vectors, finally uses SVM classifier for classification.

Related Work

VSM. Vector space model was proposed in 1975 by the famous IR scholars Salton, whose main idea is the text as a weighted vector set of words, on behalf of the importance of the weight of each word. Some symbols have specific definitions: dictionary $V=\{v_1, v_2, \ldots v_n\}$, N is the total number of words; text set $D=\{d_1, d_2, \ldots d_n\}$, M is the total number of text; a single document vector $d_i \in D$ is represented by $\{w_1^i, w_2^i, \ldots w_N^i\}$, where w_K^i represents $V_k \in V$ weight in d_i, usually used to measure the TF-IDF:

$$w_k^i = tf_{ki} \times \ln \frac{M}{tf_k} \quad (1)$$

Among them tf_{ki} represents the number v_k appears in d_i, the greater the value, the more important for the representation v_k text d_i. tf_k represents the total numbers of text containing v_k, the greater the value, the text indicates that the word v_k characteristic contribution lower for d_i.

Information Gain. Information gain is an effective feature selection method [12]. It can bring the feature as a measure of how much the importance of bringing the more information is information classification system, the more important features. For vocabulary *t* and document category *c*, calculating t for the calculation of information gain of *c*. In this paper, the following definition formula:

$$\mathrm{InfGainTxt}(t) = -\sum_i \mathrm{P}(c_i) \log \mathrm{P}(c_i) + \mathrm{P}(t) \sum_i \mathrm{P}(c_i|t) \log \mathrm{P}(c_i|t) + \mathrm{P}(\bar{t}) \sum_i \mathrm{P}(c_i|\bar{t}) \log \mathrm{P}(c_i|\bar{t}) \quad (2)$$

Where $P(t)$ represents the probability word *t* appears, with the number of documents that appeared vocabulary *t* divided by the total number of documents showing; $P(c_i)$ represents the probability c_i appears, the number of documents belonging c_i category divided by the total number of documents represents; $P(\bar{t})$ indicates that the document does not contain the word *t* is the probability, $P(c_i|t)$ contains the word indicates time *t*, c_i conditional probability of belonging to the text, also $P(c_i|\bar{t})$ indicates corpus *t* does not contain the word, the part of the text c_i probability.

The higher the value of Information gain, the greater information the vocabulary of classification system brings; The lower the value of Information gain, the fewer information the vocabulary of classification system brings. In this paper, the author use information gain to build words whitelist and the words for the calculation of the contribution of the classification system, to improve the efficiency of text classification.

Word2Vec. Word2Vec is a productivity tools that word will be represented as a real-valued vectors made by Google in mid-2013, using continuous Bag-Of-Words and Skip-Gram both models [13], As shown in Fig. 1.

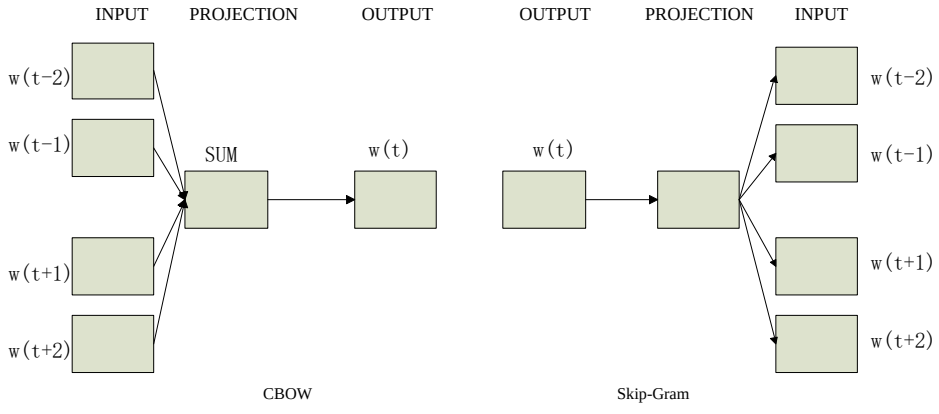

Figure 1. CBOW and Skip-Gram model

CBOW and Skip-Gram model contain an input layer, a hidden layer and output layer. Wherein, CBOW model predicts the current word by the context, Skip-Gram model predicts its context by the current word. It offers two optimization methods to improve training efficiency word vectors, respectively Hierachy Softmax and Negative Sampling. Through training corpus, each word can be mapped to the K-dimensional real vector; by the distance between the words can determine semantic similarity between them. Therefore, the word vector Word2Vec output can be used to do a lot of NLP related work, such as clustering, finding synonyms, speech analysis and so on.

Based Word2Vec Short Text Classification Feature Extension.

Short text classification includes training phase and testing phase two processes, as shown in Fig. 2. Specific process is that the author preprocesses the training data algorithm and obtains feature vectors by feature extraction algorithm the, and then expands the text feature vectors by Word2Vec; after the pretreatment of the test data set, uses Word2Vec to expand feature vector after using the training set of dictionary feature vector representation of the text, and finally classifies text using SVM.

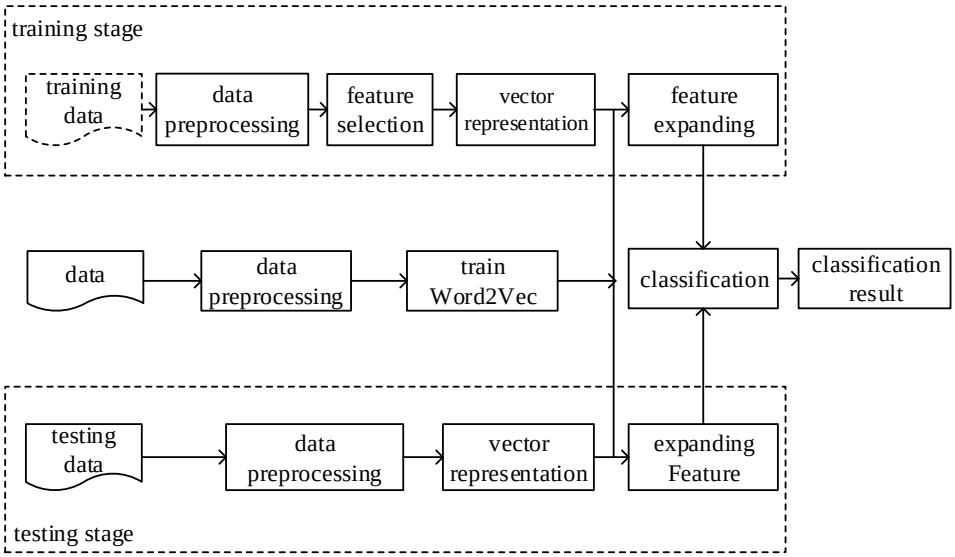

Figure 2. The short text classification flowchart base on word2vec features extension

Text Preprocessing and Feature Selection. First, the author uses Language Technology Platform to segment ate words for the training set of documents [14], followed by Chinese stop list filtering out stop words, part of speech is filtered to remove not related symbols such as @ and so certain to retain the core part of the sentence. To construct dictionary, record information for each word appears, such as the total number of words appearing in the document, the number of all the documents and categories of documents that contain the word appears. Using equation (2) to calculate its information gain value for each word, and to take the top-k words as the feature words by the value of information gain, and every word were characterized by a unique number.

For the test data set, the pretreatment is generally similar to the text on the training data set, but in the choice of features, since the text of the testing phase haven't the category label, the feature set of test data dictionary is same as the feature set of training data dictionary that the test set and the training set common feature dictionary.

Vector Representation. In order to using LIBSVM tool, the author format the documents of training set as follows: Category ID number: eigenvalues. Category ID is the category number that document belongs to; the number is a characteristic feature in the feature dictionary numbers, eigenvalues with the formula (1) values calculated. The following are examples 1 100: 0 0.0743602816966 1000: .07178746497194445 2001: 0.426938617999 ..., where 1 indicates the

category, the number 100 represents the number in features dictionary, 0.0.0743602816966 represents eigenvalues.

For a document of test set, after preprocessing, to check if the feature contains words in the dictionary, and if so, choose these words are characterized and calculate its eigenvalues calculation method or according to equation (1), which tf_{ki} represents the frequency that v_k appears in the test document, tf_k represents the total number of the training data set containing v_k . The result of text classification directly is bad due to less words and sparse features on short text, direct text classification results in general poor.

Extended Features. Because of sparse feature on short text, this paper uses feature extension method. The method is: the author trains Word2Vec model through a large set of documents, and finds out the meaning of their closest n words to a document in a word of each feature separately by using Word2Vec model, then adds to the feature vectors, as the text semantic feature extension.

To train document set that Word2Vec model requires, and crawl part of Sina microblogging data nearly a month, six topics of education, science and technology, sports, entertainment, military, economic data is selected to manually label. In order to make different categories of data more balanced, each category is about 3000.

Specific approach is to use a large set of documents to obtain the relevant training Word2Vec model parameters can simplify the process of text content for K-dimensional vector space vector operations, and the similarity vector space can be used to represent text semantically similarity. Then to find out the meaning of their closest n words to a document in a word of each feature separately by using Word2Vec model, then add them to the feature vectors, if the extended features added to the original already exists the feature, on the use of the original characteristic value, if a word appears many times in the feature extension feature in word value takes the maximum value.

Because for each characteristic word is calculated by Word2Vec cosine of its similar words and the corresponding list, and the value range is between 0 to 1, the greater the value is, the higher of these two words on behalf of the association. The authors propose the method used to calculate the eigenvalues extended feature words, if the corresponding cosine is larger the extension feature words eigenvalues is closer the expanded feature words eigenvalues, these two words is more likely to belong to the same topic. The calculation method is as follows:

$$w_{kj} = w_k \times r_{kj} \tag{3}$$

Wherein, w_k is the value of feature words v_k, r_{kj} is the value of similarity between extended feature word and feature words.

Text Categorization. In this paper, the author adopts SVM classification methods and specific tool is LIBSVM. LIBSVM an easy-to-use and fast and efficient SVM pattern recognition and regression package which designed by Zhiren Lin who is associate professor of Taiwan University [15],. One algorithm is the tool used by many types of pattern recognition and more effective solution to the problem of classification.

Experimental Results and Analysis

Experimental Corpus. The author crawled part of the Sina microblogging data through web crawler, and chose six topics such as microblogging education, science and technology, sports, entertainment, military, economic. A total of 17,760, of which 2926 Education, Science and Technology 2901, 3003 sports, entertainment 3050, 2950 military, economic 2930. For classification of short text, we use one of the 13320 as the training corpus, 2220 for each category, and the remaining 4440 title for testing corpus.

Evaluation. In order to evaluating short text classification system, we use the following evaluation criteria: accuracy P, the recall rate R, as well as the value of F1, they are defined as follows:

Accuracy:

$$P = \frac{\text{Assigned to correct certain types of text data}}{\text{The actual number assigned to certain types of text}} \quad (4)$$

Recall:

$$P = \frac{\text{Assigned to correct certain types of text data}}{\text{The actual number assigned to certain types of text}} \quad (5)$$

F1:

$$\text{F1} = \frac{2 \times P \times R}{P + R} \quad (6)$$

Experimental Results and Analysis. In this paper, the experiment were divided into three groups: the first group uses the traditional text classification methods, the feature is not extended; the second group uses the LDA classification method based on the characteristics of extension; the third group uses the feature extension classification method based on Word2Vec . In the experiment, we select the maximum 3000 words eigenvalues of the information gain uniquely numbered for each word, thus constituting features dictionary. In a second set of experiments, we chose the Chaozhen Lv's paper implementation [11]. In the training process, we choose a topic number 60, α is 0.82, β is 0.01; the number of keywords for each topic 100, the number of iterations is set to 1000. In a third set of experiments, by trial and error, adopting Skip-Gram model when training Word2Vec and mapping each word vector dimension is 200. The size of the training is setted to the window 5 (consider each of the first five words and after five words), learning rate of 0.025 to the default; in order to determine the use of extended vocabulary Word2Vec number n, set it in the range of 10 to 50 (interval to 5), through several tests, with values increases, the effect of rising, when the number increased to 30, the effect is not obvious improvement, and the computer is becoming slower these considerations that the value is set to 30. Precision and recall rate each experimental results are summarized in Table 1, below.

Table 1 High and low settings of predictor variables

category	VSM		LDA		Word2Vec	
	P	R	P	R	P	R
education	0.775	0.791	0.801	0.810	0.830	0.842
science	0.768	0.775	0.811	0.815	0.818	0.820
sports	0.813	0.805	0.848	0.835	0.852	0.849
entertainment	0.750	0.761	0.795	0.798	0.793	0.801
military	0.803	0.806	0.844	0.835	0.860	0.862
economy	0.761	0.757	0.785	0.776	0.802	0.811

The accuracy rate of more than 6 categories and averaged recall rate and average seek F1 each category, comparing the results of the three methods shown in Fig. 3.

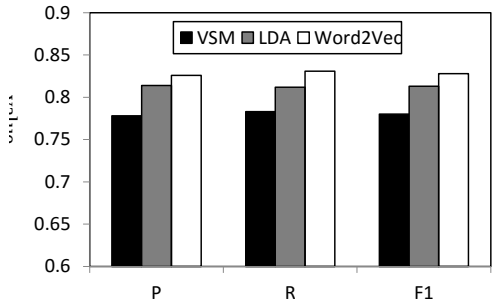

Figure 3. Comprehensive comparison of three methods

By observing the above results found, Word2Vec features scalable approach used in this paper in terms of precision and recall rates have greatly improved compared with the traditional method of VSM. With respect to the accuracy and recall rate LDA method of classification has also been a slight increase. This feature words defined by looking for Word2Vec related vocabulary, extension feature vector method can improve the short text classification results.

Through experimental results, the effect of the experimental category for "entertainment" category is significantly worse than the other categories. By querying the "entertainment" category corpus we find out that the contents of the corpus contains very broad portion of the text that contains the characteristics of the other categories, it may result in the classification results worse than the other categories.

Summary

In order to solve the problem of short text features sparse, this paper presents a method based on Word2Vec feature extension, through training a large set of documents, the original feature vectors semantic extension. The contrast experiments, the proposed method of short text classification has a good effect, therefore, the paper presented a short text Word2Vec extended feature information-based approach is feasible.

Further, since the natural language expressions as human thought is advanced, coupled with similar microblogging and other social media expressed non-standard, especially short text processing also is faced with many new challenges. For a multi-meaning and synonyms appear in the text classification problem in this study have not yet been considered in the follow-up study will consider the above scenario.

Acknowledgements

National Natural Science Foundation of China (Grant No. 61272427)

References

[1] C.H. Li: *The 15th National Computer Science & Computer Continuing Education Conference* (*Qinhuangdao, China*, 2004).Vol. 1, p.7.

[2] D.H. Fang:*The Research and Implementation of Microblogging Short Text Classification Based on LDA*(MS., Northeastern University,China 2011),p.26.

[3] M. Sahami, T.D. Heilman: *Proceedings of the 15th Conference on World Wide Web* (New York,U.S, 2006)Vol.1,p.377-386.

[4] W. Yih, C. Meek: *Proceedings of the 22nd Conference on Artificial Intelligence*.(Menlo Park,U.S, 2007)Vol.1,p.1489-1494.

[5] Y.D. Zhai, K.P. Wang and D.N. Zhang: An Algorithm for Semantic Similarity of Short Text Based on, Vol. 40 (2012) No.3, p.617-620.

[6] S. Banerjee, K. , *Gupta: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007: 787-788.

[7] B. Sriram, D. Fuhry: Short Text Classification in Twitter to Improve Information Filtering [A], SIGIR'10[C]: 841-842

[8] Danesh Irani, Steve Webb. Study of Trend-Stuffing on Twitter through Text Classification[A],CEAS 2010[C],July 13-14: 114-123

[9] Y.H., Q. Zhu: Semi-supervised short text categorization based on attribute selection, Vol. 30 (2010) No.4, p.103-111.

[10] X.W. Wang, X.H. Fan: Method for Chinese short text classification based on feature extension, Vol. 29 (2009) No.3, p.108-199.

[11] C.Z. Lv, D.H. Ji and F.F. Wu: Short text classification based on expanding feature of LDA,Vol. 51 (2015) No.4, p.123-127.

[12] D.H. Fan: *Research for feature selection algorithm based on text clustering* (Normal University,China 2012),p.36.

[13] J. Zhang, D. Qu and Z. Li: Recurrent neural network language model based on word vector features, Vol.4 (2015), p.299-305.

[14] T. Liu, W.X. Che and Z.H. Li: Language Technology Platform, Vol.6 (2011), p.53-62.

[15] Z.X. Yang, J.F. Liu and L. Sun: Using invariants to predict the potential for errors in programs, Vol. 4 (2010) ,p.327-331.